Journal of Quantitative Criminology, Vol. 9, No. 2, 1993

Minimizing the Cost of Dispatch Delays by Holding Patrol Cars in Reserve

Stephen R. Sacks,[1] Richard C. Larson,[2] and Christian Schaack[3]

At many police departments high-priority callers sometimes incur undue delays that could be avoided by the use of a differential response strategy that takes full account of the different "costs" of delay for different priority calls. In this article we examine such a strategy, in which lower-priority callers may not be served immediately, even though some response units are available. Arriving priority i callers are queued whenever the number of busy patrol units equals or exceeds the cutoff number for that priority. Our purpose is (1) to find a practical way of choosing the set of cutoff numbers that will minimize the expected total cost of delays for the entire system and (2) to use that method to investigate how the optimal set of cutoffs changes in response to changes in several important variables, including the relative costs of delay for different priorities, the overall workload, and the relative frequencies of different priority calls.

KEY WORDS: differential response; police deployment; police dispatching; 911.

1. INTRODUCTION

Efficient management of police patrol resources requires a dispatch policy that facilitates immediate response to high-priority, perhaps life-threatening, calls for service. Although most police departments have a set of categories that are used to classify calls by priority group, in fact such priorities usually play only a limited role in dispatching police response units: when a call of any priority is received, most departments immediately send a car if at least one is available and place callers in the dispatcher's

[1]Department of Economics, University of Connecticut, Storrs, Connecticut 06269.
[2]Operations Research Center, MIT, Cambridge, Massachusetts.
[3]**Banque Generale du Luxembourg.**

0748-4518/93/0600-0203$07.00/0

queue only if all patrol units are busy. Often the priority assignments are used only to select calls from the dispatcher's queue, with higher-priority calls given greater preference for being pulled first. Consequently, high-priority callers sometimes incur undue response delays that could be avoided by a dispatching system that recognizes the different "costs" of delay for different priority calls and sometimes deliberately delays response to lower-priority calls even when patrol units are available.

In this article we examine a differential response strategy that improves the probability that response units will be available for high-priority callers. Lower-priority callers may not be served immediately, even though a car is available. Arriving priority i callers are queued whenever the number of busy patrol units equals or exceeds the priority i cutoff number, N_i. A queued priority i caller will be served when fewer than N_i units are busy and all higher-priority queues are empty. Worded differently, if N is the total number of cars, then $N-N_i$ is the number of cars held in reserve when dealing with priority i callers. Clearly, the cutoff for the highest priority class, N_1, will be equal to N; that is, for the highest priority calls, the only situation in which we will not send a car is when *every* car is busy. The cutoff for each lower priority is less than or equal to the previous one ($N_3 \leq N_2 \leq N_1$), and for the lowest priority, N_i may even be zero; that is, in certain circumstances it may make sense to refer the lowest-priority callers to some alternative agency and not send a response unit even if all cars are available for assignment.

Our goal in this research was (1) to find a practical way of choosing the set of cutoffs that will minimize the expected total cost of delays for the entire system and (2) to use that method to investigate how the optimal set of cutoffs changes in response to changes in several important variables, including the relative costs of delay for different priorities, the overall workload, and the relative frequencies of different-priority calls.

Clearly, the total cost of delays for the entire system depends on the cutoffs chosen for the priority classes. To find an efficient overall allocation of a given number of vehicles, it is necessary both to calculate the expected waiting time for callers in each priority class and to assign a cost to each minute's delay for each priority class. Then it is possible to calculate the expected cost across all priority classes, with each class weighted by the relative frequency of its calls and its assigned cost, to get a "weighted wait," which we call the Expected Total Cost of Delays.[4] If this is done for each

[4]The Expected Total Cost of Delays may be expressed as $\sum_{i=1}^{j} \lambda_i/\lambda_T * W_i * C_i$, where j is the number of priority classes, λ_i is the call for service rate for priority i, λ_T is the total call rate, W_i is the expected wait (in minutes) for priority i callers, and C_i is the cost (in dollars or in any other measure of seriousness) of a 1-min wait for priority i callers.

of several different sets of cutoffs, we can find the set that has the lowest expected total cost.

A difficulty with this approach is that it is not obvious in advance how to choose sets of cutoffs for which to do the calculations and comparison. The number of possible sets may be very large and each one requires a substantial amount of calculation. (If 12 cars are available and there are 4 priority classes, the number of possible sets is 364; with 5 classes the number of possibilities is 1365.) To avoid having to do the calculations for all feasible sets of cutoffs, we have developed a gradient search method which allows us to calculate and compare the weighted wait for only those sets of cutoffs which are likely to reduce the value of our objective function.

Throughout this article we deal with systems that have three priority classes, but in principle there could be any number.

2. RELEVANT LITERATURE

Use of operations research methods, including the queueing theory used extensively in this article, is now a widespread, widely accepted practice in criminal justice. The scientific study of queues started in 1915 in Denmark when the Danish telephone engineer A. K. Erlang developed the first mathematical models of a queue to assist the Copenhagen Telephone Company correctly "size" its first telephone switching systems. Like other components of operations research, queueing theory helps us understand a complex operational phenomenon by constructing an explanatory mathematical model.

While details differ, all queues share a similar generic structure. The input to a queueing system is a stream of arriving customers desiring service. Once at a service facility, an arriving customer may enter service immediately or, if all servers are busy, the customer joins a queue with other waiting customers. Eventually, service is provided to the customer, after which the customer departs from the system. In criminal justice, the customers may be callers to 911, burglar or robbery alarms to a central station, or prisoners queueing up for processing. The servers are police officers, desk clerks, judges, assistant district attorneys, etc. If the arriving customer is delayed in queue, she is not necessarily selected for service from the queue in a first-come, first-served manner; in criminal justice settings, queued "customers" are often assigned different priorities, with higher-priority customers being served from the queue before lower-priority customers. A 911 caller reporting a current break-in to her home would, for instance, receive priority over a report of a stolen TV set from the rear seat of an automobile.

The mathematical theory of queues has been applied to problems of police dispatch and deployment since the early 1970s (Bottoms *et al.* 1972;

Larson, 1972a, b; Larson and Chaiken, 1972). With the well-known PCAM (Patrol Car Allocation Model) program (Chaiken, 1978; Chaiken and Dormont, 1975; Chaiken and Dormont, 1978; Larson and McEwen, 1974), queueing models have been incorporated into broader police patrol deployment models that include not only dispatcher queueing, but also preventive patrol intensities, workloads of the patrol units, travel times, crime coverage rates, and other factors. The purpose in each case is to help police planners schedule and distribute scarce patrol units across commands (e.g., precincts) in an optimal manner.

The unique problems of municipal police departments have created queueing situations previously unseen by queue theorists and, thus, have helped spur new and exciting research results in the mathematics of queues. One example is the Hypercube Queueing Model (Larson, 1974a, b, 1975), which incorporates a region's geography and call-for-service patterns to model the behavior of a proposed spatial deployment of police patrol resources. The intent is to help the police planner balance the conflicting objectives of minimizing overall average response time, minimizing differences in neighborhood-specific average response times, and minimizing differences in the workloads of the individual patrol units. Since its creation in 1974, this model has been substantially generalized to include such factors as automatic vehicle location dispatch (Larson and Franck, 1978), patrol-initiated activities (Larson and McKnew, 1982), and dispatch of multiple units (Chelst and Barlach, 1981). Other recent advances include more realistic travel time estimates and computer-generated optimal patrol beats (Larson, 1989). The model has now been adapted to run on a desktop computer and has been used by police departments in Orlando, Florida, Chapel Hill, North Carolina, and Hartford, Connecticut.

In the late 1970s the National Institute of Justice (NIJ) of the U.S. Department of Justice funded a study of police response and deployment in the Wilmington, Delaware, Police Department (Cahn and Tien, 1981). The key recommendation of the empirically based study was to implement a radically new police dispatch policy that explicitly and aggressively managed the calls for service that arrived at the police dispatcher's position. This strategy, under the name "police differential response strategy," was trial tested by three other police departments, with support of the NIJ (McEwen *et al.*, 1986). The trial tests were judged successful, and police differential response strategies were then advocated nationally by the NIJ. For a full explanation of police differential response strategies, see McEwen *et al.* (1986); for a brief review focused on the queueing and related operational aspects, see Larson (1990) or Schaack and Larson (1989).

A key element of police differential response strategies is the deliberate delay at the dispatcher's position of nonurgent calls for service, even if patrol

units are available, in order to keep those units in "tactical reserve ready status" for near-term urgent calls that may or may not arrive. Over the long haul, such a strategy reduces the average cost of police response delay, for reasonable definitions of cost of delay. The growing importance of this policy, which represents a sharp shift from more traditional "taxicab-oriented police dispatching," led to important new results in the mathematics of queues, queues that incorporated such deliberate delayed response (Schaack and Larson, 1986) and queues that also incorporated dispatch of multiple police patrol units (Green, 1984; Green and Kolesar, 1989; Schaack and Larson, 1989). The model that we utilize in this paper is perhaps the most advanced of these, incorporating both deliberate delayed response and dispatch of multiple units (Schaack and Larson, 1989).

3. THE DATA

Ideally, calculation of optimal response delays would be done using complete, recent data on number of calls for service by priority class, along with exact information on number of vehicles dispatched and service time, for each hour of every day. To avoid basing policy on atypical situations, the data should be averaged over several weeks. For this study we did not have such ideal data. All of the empirical work was done twice, once using data provided by the Police Department in Peoria, Illinois, a city with a population of 125,000, and then again using data provided by the Police Department in Hartford, Connecticut, a city of 140,000.[5] For Peoria we have the number of calls for service for each hour of each day of the week from June to August of 1986. For Hartford we have the number of calls for each hour of the first week in January of 1991.

The 1986 Peoria data are not broken down by priority class. However, for June 1985 we do have hourly totals for each of three priority classes in Peoria. We assume that the distribution of calls across priority classes did not change from 1985 to 1986 and, further, that these proportions are the same for every day of the week. This allows us to calculate for each hour of the week the proportion of the total number of calls in each priority class. We then multiply all of the June–August 1986 data by these proportions to get the number of calls per hour for each priority class for each hour of each day (and for a day we call "ALL," i.e., an average day). These are the calls-for-service rates that we use in our Peoria tests. The Hartford data are more complete: we have the number of calls from each priority class for every

[5]We are grateful to Aubrey Moore of the Peoria Police Department and to James Donnelly of the Hartford Police Department, without whom this research would not have been possible.

hour. Hence we do not need to make assumptions about the distribution of calls across priority classes.

For both cities our data on service time per call are not broken down by day of the week, so we assumed that there is no difference between days, but the data do vary by time of day. Thus our results are based on the assumption that service time (μ) varies by hour but not by day or by priority class.

In our calculations we use the probability of a priority i caller requesting n servers [$P(i, n)$]. For Hartford we know for every hour of the week, and for each priority, the number of calls that required one or more than one car. (We assume that more than one means exactly two.) But for Peoria we have only the proportion of priority one calls to which one car was dispatched, and that we have for only certain time blocks during May and June of 1983. With some reasonable assumptions we generated a complete set of $P(i, 1)$ and $P(i, 2)$ for Peoria, but clearly the Peoria data are less reliable than the Hartford data. For that reason, most of the specific numerical results presented in this article are based on the Hartford data. But we would like to emphasize that what we found when we redid all of our calculations using Hartford rather than Peoria data was that all of our initial conclusions, without exception, were confirmed. Thus, while the limitations of our data may reduce the accuracy of some specific results (i.e., how many cars to hold in reserve at a particular hour), we are confident of the general conclusions we present below.

4. THE COMPUTATIONAL PROCEDURE

The heart of the calculations that must be done is not new. Schaack and Larson (1989) describe a procedure for calculating the expected wait for each priority class in a multiserver queuing system. What we are adding here is a procedure for repeatedly making those calculations, each time with a different set of cutoffs, in a sequence that allows us to find the optimal set. Then we can investigate how that optimum is affected by variations in the relative costs of delays and in the calls-for-service rates.

In order to avoid calculating the expected cost for every feasible set of cutoffs, we have developed a gradient search method that begins with all cutoffs equal to the total number of vehicles (which is equivalent to having no cutoffs) and then systematically reduces the cutoffs for all but the top class. Figure 1 shows an example with three classes. We can think of this as moving through two-dimensional space: the cutoff for priority class one remains fixed at N (the total number of vehicles), the vertical coordinate measures the cutoff for priority class 2, and the horizontal coordinate measures the cutoff for priority class 3. We begin at point (N, N) and move

cutoffs			Expected Wait		Improvement
N_1	N_2	N_3	previous	current	
14	14	14		14.70	
14	13	13	14.70	10.20	4.55
14	12	12	10.20	4.77	5.42
14	11	11	4.77	3.25	1.52
14	10	10	3.25	2.40	.85
14	9	9	2.40	2.87	-.47
14	11	10	2.40	2.14	.26
14	12	10	2.14	2.27	-.09
14	11	9	2.14	1.73	.41
14	11	8	1.73	9.76	-8.03
14	12	9	1.73	1.78	-.05
14	10	9	1.73	1.93	-.20
optimum is 14	11	9		1.73	

Number of iterations: 12 Number of feasible sets of cutoffs: 61

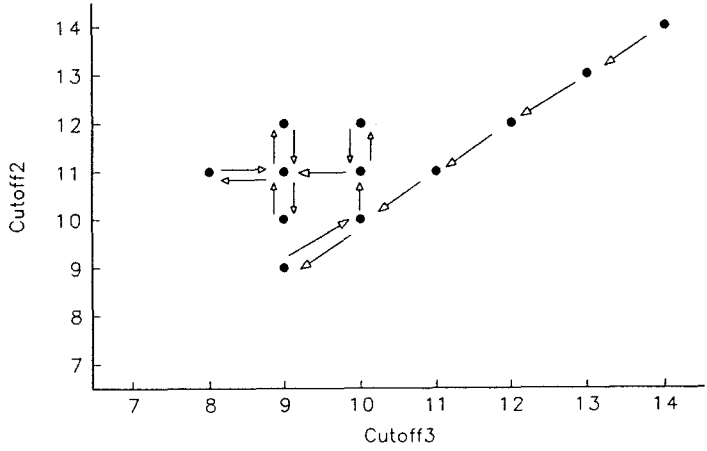

Fig. 1. Example of finding optimal cutoffs with the gradient method. Arrows indicate the path taken by the gradient method in its search for the optimal set of cutoffs.

down the diagonal (i.e., reduce by one the cutoffs for priority classes 2 and 3) as long as each step improves the objective function. That is, at each step we calculate the Expected Total Cost of Delays for the current set of cutoffs. Once that leads to a negative improvement, the program backs up one step and tries one-unit changes in the cutoffs, one at a time, continuing in any direction as long as improvements are positive, and concluding that it has found the optimum when no improvement is found in any direction.

We are relying here on an assumption that the cost function has a local minimum, even though we have not formally proved convexity. This assumption is supported by examination of the data, which shows that as

we vary one coordinate, holding the others constant, the size of the improvement diminishes as we approach the optimum. Furthermore, for dozens of cases we have plotted the Expected Total Cost vertically (the Z axis) and the priority 2 and 3 cutoffs horizontally (the X and Y axes). Figures 2 and 3 show several examples. (The function is defined only at the lattice points, but the figures show a surface so as to make the relationship easier to see.) Clearly, the underlying relationship appears to be convex.

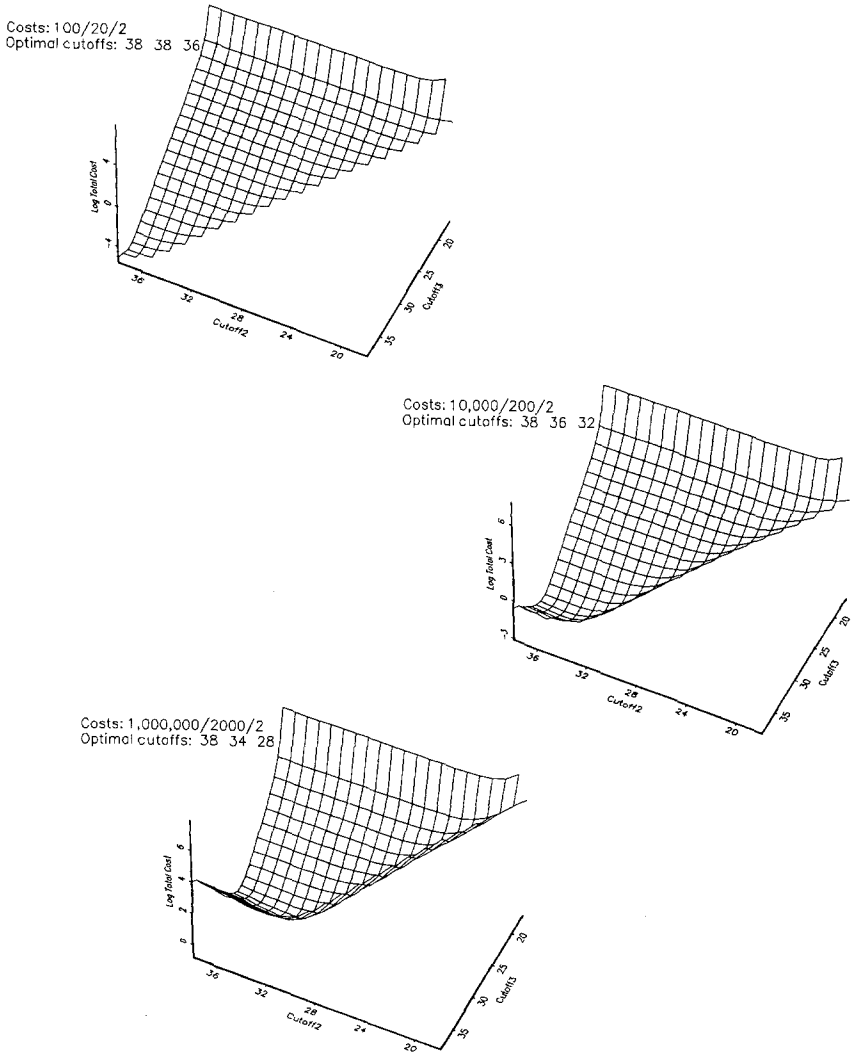

Fig. 2. Expected total cost of delays, Monday at noon.

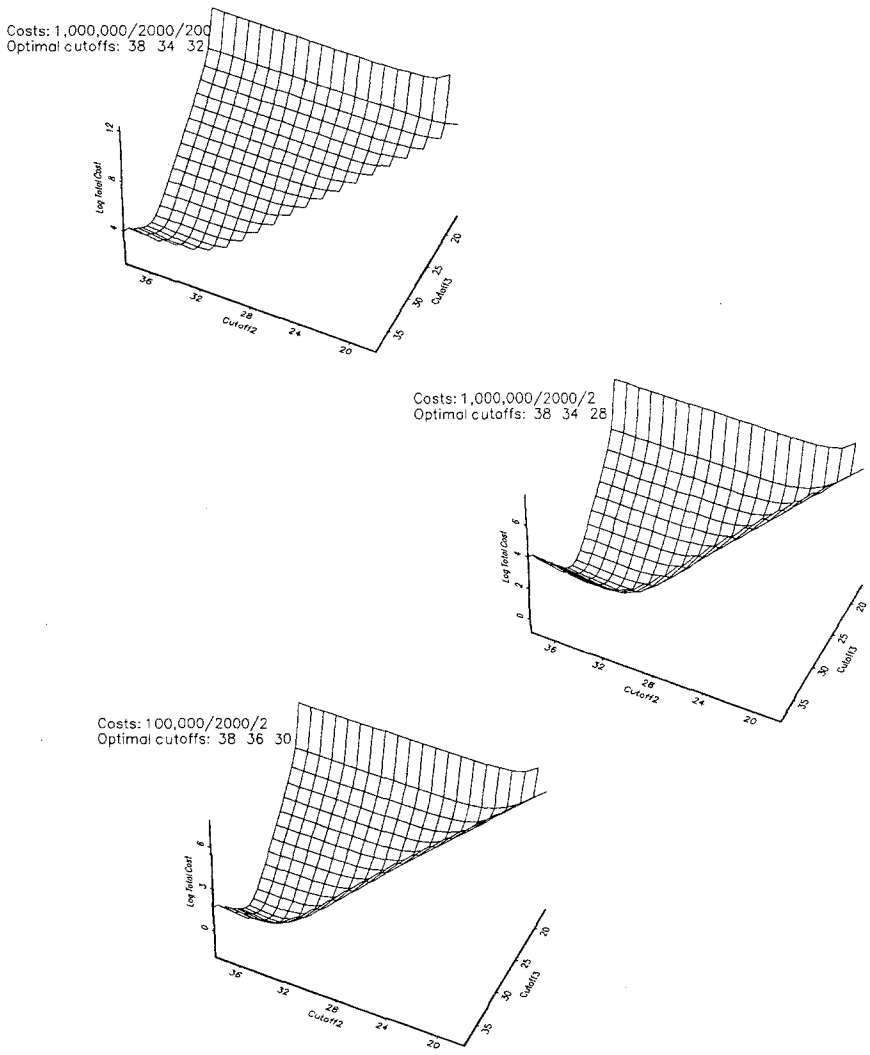

Fig. 3. Expected total cost of delays, Monday at noon.

We used this method to find the optimal cutoffs for our test data, with various sets of costs. For every fourth hour of the entire week, including an eighth day called ALL, which is an average of the other days, we found the optimal set of cutoffs using first this gradient method and then using exhaustive enumeration. In all 48 of these tests the two methods found exactly the same optimum. This 100% success rate in tests, using both Peoria and Hartford data, lends considerable support to the gradient search approach.

Table I: Comparison of the Numbers of Steps Required to Find Optimal Set of Cutoffs[a]

Day	Time	CFS/hr	Number of cars	S1	S2
			Hartford		
Mon	0	32.0	32	89	8
Mon	4	11.0	21	51	10
Mon	8	13.0	21	76	10
Mon	12	21.0	38	251	11
Mon	16	34.0	38	152	10
Mon	20	25.0	32	118	10
Tue	0	21.0	32	189	10
			Peoria		
Wed	21	16.8	16	34	8
Thu	2	9.5	16	89	8
Thu	6	4.4	11	55	7
Thu	10	8.8	12	36	7
Thu	14	11.8	12	26	8
Thu	18	12.2	11	20	7
Thu	22	20.8	16	28	8
Fri	2	9.8	16	78	10

[a] S1 is number of feasible sets of cutoffs. S2 is number of steps taken by gradient method.

Table I compares the number of steps required by the two methods to find the optimum for a representative set of tests. For instance, with Hartford data for Monday at midnight exhaustive enumeration required calculating the weighted expected wait 89 times and the gradient method found the same optimum with only 8 such calculations. Note that 89 is the number of *feasible* sets of cutoffs; many more could be specified, but for them the arrival rate for one or more categories would exceed saturation. The number under S1 does not include infeasible sets. Hence the advantage of the gradient method, although still substantial, is less in this case than for, say, Monday at noon, where using the gradient method reduces the number of steps from 251 to 11.

The tests were repeated using heavier weights (costs) for higher priority callers and examining every twelfth hour of the week: in all of these 16 tests the gradient method found the same optimal cutoffs as did exhaustive enumeration.

5. EMPIRICAL RESULTS

5.1. The Amount of Cost Reduction

The central point of this article is that judicious use of priority queues with cutoffs can reduce the cost of delays. Table II makes this point very

Table II. Costs of Delay for a Random Caller With and Without Cutoffs for a Selected Sample of 31 Hours (Sorted by Total Number of Vehicles Available) (Costs: 5000/200/20)

Day	Time	Calls for service per hr	Expected cost without cutoffs	Expected cost with cutoffs	Cost reduction (%)	Optimal cutoff N_1	Optimal cutoff N_2	Optimal cutoff N_3
Mon	5	16.0	667.87	212.45	68	21	19	18
Tue	6	13.0	1.84	0.95	48	21	21	19
Wed	2	11.0	3.08	2.96	4	21	21	20
Wed	7	19.0	57.17	21.80	62	21	19	18
Fri	4	4.0	0.02	0.01	50	21	19	18
Fri	9	18.0	46.30	15.40	67	21	20	18
Sat	5	13.0	0.86	0.86	0	21	21	21
Sun	6	9.0	0.00	0.00	0	21	21	21
Mon	0	32.0	10.95	5.13	53	32	32	30
Mon	20	25.0	8.42	4.53	46	32	31	29
Tue	1	12.0	0.00	0.00	0	32	32	29
Tue	21	28.0	1766.08	976.09	45	32	30	29
Wed	22	28.0	876.99	403.22	54	32	31	30
Thu	18	36.0	319.77	116.14	64	32	30	29
Thu	23	27.0	174.59	65.42	63	32	30	28
Fri	19	29.0	176.14	110.79	37	32	32	29
Sat	0	26.0	0.47	0.15	68	32	31	29
Sat	20	26.0	12.74	9.06	29	32	32	29
Sun	1	24.0	42.27	12.99	69	32	30	28
Mon	10	16.0	0.09	0.03	67	38	36	34
Mon	15	25.0	0.04	0.02	50	38	38	36
Tue	11	27.0	3.34	1.67	50	38	37	34
Tue	16	42.0	8.47	5.65	33	38	38	36
Wed	12	23.0	0.10	0.04	60	38	37	35
Wed	17	29.0	0.83	0.39	53	38	37	35
Thu	13	37.0	20.15	14.53	28	38	38	36
Fri	14	24.0	1.21	0.56	54	38	36	34
Sat	10	29.0	34.11	16.26	52	38	37	35
Sat	15	28.0	0.21	0.09	57	38	37	35
Sun	11	26.0	0.45	0.31	31	38	38	36
Sun	16	25.0	0.05	0.01	80	38	36	34
Average					47			

clearly. For every fifth hour of the week (omitting two for which there were data errors), we have calculated the expected cost of delay twice, once using the optimal set of cutoffs as determined by the gradient search method and once without using cutoffs (that is, a response unit would be dispatched if any were free, regardless of the priority of the call).[6] The expected cost for

[6]To facilitate comparison among times at which the same number of vehicles are available, we have sorted the results by the priority 1 cutoff, which is always equal to the total number of vehicles.

each of the three priority categories is calculated by multiplying its expected wait in minutes by its dollar cost per minute; then the overall expected cost is found by taking a weighted average of these three, where the weights are each category's share of calls. The result is the expected cost of delay for a random caller. Alternatively, one might multiply each of the three expected costs by the number of calls for that priority (rather than the proportion of calls) and then sum, getting the total cost to all callers, rather than the expected cost for a random caller. Using this alternative calculation would not change the optimal set of cutoffs or the percentage reduction in cost, but the magnitude of the reduction would be larger by a factor equal to the total number of calls for that hour.

For example, for Monday at 5 AM our gradient method found the optimal set of cutoffs to be 21/19/18, giving expected waiting times of 0.10, 1.32, and 5.83 min, respectively, for the three priority categories. Multiplying these three waits by costs of \$5000, \$200, and \$20/min respectively, we get the expected costs of waiting for the three categories: \$500, \$264, and \$116.60. At that hour the number of calls for the three categories are 4, 0, and 12, so we multiply \$500, \$264, and \$116.60 by 0.25, 0.0, and 0.75, respectively, and sum to get the overall expected cost of waiting, \$212.45. If no cutoffs were used, the three expected waits would be different (longer waits for the higher priorities) and the overall expected cost would work out to \$667.87. Thus the differential response strategy reduces expected cost by 68%. Of course, these numbers depend on the arbitrarily assigned costs of waiting, as well as on the relative frequency of each priority's calls and the total number of cars available. In Section 3.2 below we discuss the effect of assuming a different set of costs. If, for Monday at 5 AM, we were to multiply the expected cost for each of the categories by 4, 0, and 12, instead of by 0.25, 0.0, and 0.75, the overall expected cost would be 16 times larger (16 is the total number of calls). Consequently, the cost reduction attributable to the use of cutoffs would also be 16 times larger: \$7286.72 (=\$10,685.92 − \$3399.20). In terms of percentages, the reduction would still be 68%. For the particular set of costs used in Table II, the average savings for the 31 hr examined is 47%.

It is important for police planners to recognize that the optimal set of cutoffs is different for different times of day and different days of the week. This, too, is evident in Table II, which shows that for a given number of available vehicles, there is not a unique set of optimal cutoffs. In the remainder of this article we report on our investigation of the effect on the optimal cutoffs of changes in each of several important variables, *viz.*, the relative costs of delays, the overall workload, and relative workloads (e.g., an increase in the number of priority 1 calls relative to priority 2 and priority 3 calls).

5.2. The Relative Costs of Delay

We used an initial simulation to investigate the effect of varying the relative costs of a minute's delay. We began by choosing, for each of the three priority categories, a number that represents the cost (in dollars or some arbitrary units that measure seriousness) of a 1-min delay in responding to a call. Since the costs are used as weights, only their relative magnitudes are significant. Obviously, the cost will be higher in priority 1 than in priority 2, and higher in priority 2 than in priority 3. By repeatedly calculating the optimal set of cutoffs, each time changing one of the costs, we can observe how the latter affects the former.

Table III shows the results of two sets of simulations. In the top part of the table we hold the costs (which may be thought of as dollars) of a 1-min wait constant for priorities 2 and 3 and we gradually increase the cost of waiting in priority 1. Except for the first row, which is an arbitrary starting

Table III. Cost Differentials Big Enough to Change the Optimal Set of Cutoffs (Thursday at 11 PM)[a]

Calls For Service by priority: 5.00		7.00	15.00			
Cost			Minimum expected cost of delays	Optimal cutoff		
Pr. 1	Pr. 2	Pr. 3		N_1	N_2	N_3
Cost for priority 1 incremented by 50.00 up to max. of 300,000.00						
500.0	200.0	20.0	29.166	32	32	30
1,200.0	200.0	20.0	40.152	32	32	29
1,800.0	200.0	20.0	47.054	32	31	29
3,300.0	200.0	20.0	59.207	32	31	28
3,750.0	200.0	20.0	61.606	32	30	28
16,250.0	200.0	20.0	99.780	32	30	27
18,650.0	200.0	20.0	104.988	32	29	27
51,050.0	200.0	20.0	150.707	32	28	27
72,800.0	200.0	20.0	167.801	32	28	26
209,850.0	200.0	20.0	246.879	32	27	26
Cost for priority 2 incremented by 10.00 up to max of 18,650.00						
18,650.0	30.0	20.0	75.374	32	28	28
18,650.0	50.0	20.0	81.368	32	29	28
18,650.0	180.0	20.0	102.721	32	29	27
18,650.0	210.0	20.0	105.592	32	30	27
18,650.0	1,170.0	20.0	160.059	32	31	27
18,650.0	1,210.0	20.0	161.142	32	31	26
18,650.0	2,370.0	20.0	184.539	32	32	26
18,650.0	12,330.0	20.0	283.242	32	32	25

[a]CFS are calls per hour, but costs are per minute.

point, the table shows priority 1 costs that are just big enough to change the optimal cutoffs. The top row shows that with costs of 500, 200, and 20, it is optimal to have a cutoff less than N only for priority 3 callers; that is, if even 1 of the 32 servers is free, it should be dispatched to any priority 1 or 2 caller, and if more than 2 are free (i.e., fewer than 30 are busy), a car will be sent to a priority 3 caller. When the cost of waiting reaches 1200 for priority 1, the optimal cutoff for priority 3 drops to 29; that is, if 29 servers are busy, a priority 3 caller should not be served. Clearly, as we read down Table III there is a dramatic increase in the size of the increment in priority 1 cost needed to change the optimal set of cutoffs. Indeed, from the bottom two rows we see that priority 1 waiting cost must rise from 72,800 to 209,850 in order to change the optimal cutoffs from 32/28/26 to 32/27/26. These increments do not increase uniformly: smaller changes affect the priority 2 cutoff and larger changes are needed to affect the priority 3 cutoff.

In the bottom of Table III we hold the priority 1 cost fixed at an arbitrarily picked value and show the results of increasing the cost of priority 2 waiting time. Initially, adding 20 to the cost of each minute's wait (which nearly doubles the ratio of priority two to priority three cost) is sufficient to change the optimal set of cutoffs from 32/28/28 to 32/29/28. Then larger increases are necessary to cause a change. Again, the increments do not increase uniformly.

Since one of the most troublesome aspects of applying this research to actual police operations is deciding the costs, it is reassuring to find that the optimum set of cutoffs is in many situations not very sensitive to these parameters.

In order to provide an intuitive feeling for the relationship among these variables, we present Fig. 2, which shows (for a different set of data) how the expected total cost of delays varies as the optimal cutoffs change for each of three different sets of costs.[7] All three panels in Fig. 2 refer to Monday at noon; the only difference among them is that different priority costs were used to calculate the expected total cost. Clearly, as we increase relative costs from 100/20/2 (for priorities 1, 2, and 3, respectively) to 10,000/200/2 and then to 1,000,000/2000/2, not only do the optimal cutoffs for priorities 2 and 3 decrease, but also the convexity of the curvature of the

[7]In order to compress the range of values on the vertical axis, we take the log of the objective function. This does not necessarily preserve convexity, but it gives a good idea of the shape of the underlying relationship. (If the log form is convex, then the underlying relationship is convex, but not vice versa.) Actually, total cost as a function of a set of cutoffs exists only at the lattice points where the cutoffs have integer values, but the figures show a surface so that the relationship is easier to see. The hour used for this figure was chosen because its low workload relative to the number of cars on duty results in a large number of feasible sets of cutoffs.

surface increases. That is, as the cost differentials increase, (1) it becomes efficient to hold in reserve more cars for higher-priority calls, and (2) the penalty incurred by using nonoptimal cutoffs increases. This is evident in the fact that as we look from one panel to the next, the lowest point on the surface moves to the right and away from the front of the horizontal plane (N_2 and N_3 move from 38 and 36, respectively, to 36 and 32 and then to 34 and 28), and the height differential between the lowest point and nearby points becomes greater.

Like Fig. 2, Fig. 3 allows us to see the effect of changing relative costs. The middle panel in Fig. 3 is the same as the bottom panel in Fig. 2. Comparing the middle and top panels in Fig. 3, we see that if we hold constant the costs of delay for priorities 1 and 2 and raise the cost of delay for priority 3, the optimum moves forward (toward higher cutoffs for priority 3) and the surface flattens out with respect to the cutoff 2 axis; i.e., total cost is less sensitive to changes in the cutoff for priority 2. Comparing the middle and bottom panels, we see that holding constant the costs of delay for the second and third priorities and lowering the cost of delay for top-priority calls moves the optimum leftward and forward (to higher cutoffs for priorities 2 and 3) and the surface above the left-front corner of the horizontal plane unfolds; that is, giving high (above optimal) cutoffs to priorities 2 and 3 becomes less expensive.

5.3. Changes in Overall Workload

In order to investigate how the optimal set of cutoffs is affected by changes in the overall workload, we calculated the optimal set repeatedly, first using the actual Hartford arrival data multiplied by a factor of 0.1, then multiplied by 0.2, 0.4, 0.6, etc. Table IV shows the results for three different hours. It is clear that as the overall workload increases from one-tenth of its actual value to more than twice its actual value, the severity of the optimal cutoff differentials [as measured by an index equal to $(N_1 - N_2) + (N_1 - N_3)$] increases and then decreases. That is, the optimal set of cutoffs is not very different from no cutoffs at all when the workload is either very light or very heavy. Only at moderate workloads will a differential response strategy significantly reduce expected total cost. The results for other hours and other days of the week are similar.

5.4. The Effects of Changing Relative Workloads

The classification rules that assign incoming calls to priority categories are not necessarily fixed; by changing these rules, police planners can change the relative frequency of calls for the various priorities. For example, domestic dispute could be moved from priority 2 to priority 3, or breaking and

Table IV. Effect on Optimal Set of Cutoffs of Changes in Total Workload

Day	Time	Scale factor	Effective workload (CFS/hr)	Optimal cutoff			Severity index, $(N_1 - N_2) + (N_1 - N_3)$
				N_1	N_2	N_3	
Tue	12	0.1	1.9	38	38	37	1
Tue	12	0.2	3.8	38	38	37	1
Tue	12	0.4	7.6	38	38	37	1
Tue	12	0.6	11.4	38	38	36	2
Tue	12	0.8	15.2	38	38	36	2
Tue	12[a]	1.0	19.0	38	38	35	3
Tue	12	1.2	22.8	38	38	35	3
Tue	12	1.4	26.6	38	37	35	4
Tue	12	1.6	30.4	38	37	35	4
Tue	12	1.8	34.2	38	36	34	6
Tue	12	2.0	38.0	38	36	35	5
Tue	12	2.2	41.8	38	36	35	5
Tue	12	2.4	45.6	38	37	36	3
Tue	12	2.6	49.4	38	38	38	0
Thu	4	0.1	0.6	21	21	21	0
Thu	4	0.2	1.2	21	21	19	2
Thu	4	0.4	2.4	21	20	19	3
Thu	4	0.6	3.6	21	20	18	4
Thu	4	0.8	4.8	21	19	18	5
Thu	4[a]	1.0	6.0	21	19	17	6
Thu	4	1.2	7.2	21	19	17	6
Thu	4	1.4	8.4	21	19	17	6
Thu	4	1.6	9.6	21	18	16	8
Thu	4	1.8	10.8	21	18	16	8
Thu	4	2.0	12.0	21	18	16	8
Thu	4	2.2	13.2	21	18	16	8
Thu	4	2.4	14.4	21	18	16	8
Thu	4	2.6	15.6	21	18	16	8
Thu	4	2.8	16.8	21	18	17	7
Thu	4	3.0	18.0	**21**	18	**18**	**6**
Thu	4	3.2	19.2	**21**	19	**19**	**4**
Thu	4	3.4	20.4	21	20	20	2
Thu	4	3.6	21.6	No feasible cutoffs			
Sat	22	0.1	2.0	32	32	32	0
Sat	22	0.2	4.0	32	31	30	3
Sat	22	0.4	8.0	32	31	29	4
Sat	22	0.6	12.0	32	31	29	4
Sat	22	0.8	16.0	32	30	28	6
Sat	22[a]	1.0	20.0	32	30	27	7
Sat	22	1.2	24.0	32	30	27	7
Sat	22	1.4	28.0	32	30	27	7
Sat	22	1.6	32.0	32	30	29	5
Sat	22	1.8	36.0	32	32	32	0
Sat	22	2.0	40.0	No feasible cutoffs			

[a]True workload.

entering could be moved from priority 2 to priority 1. Clearly, the planners will want to know what the effect of such changes will be on expected waits for each of the priority categories and on the expected total cost of delays for the whole system.

We ran a number of simulations in order to observe these effects. In some we initially set the probability of requesting two servers to be the same for all priority classes. In those cases moving some calls from one priority class to another does not alter the total number of requests for two servers and hence does not alter the overall workload on the system. In other cases we look at the same issue without assuming uniform probabilities of requesting two cars.

In all of these simulations the optimizing procedure was used to calculate, for each of 18 different hours of the week, the expected wait for each of the three categories and the expected overall cost (i.e., the weighted wait). The 18 hr that were examined in this way were at various times on 7 different days. Then the initial data were modified by moving first 10%, and then larger percentages, of the priority 2 callers to priority 1; after each transformation of the data, the optimization procedure was run again for the same 18 hr. Table V shows that comparison for 1 of those 18 hr. Note that for each hour, the total number of calls for service is not changed; only the relative workloads are different. The set of cutoffs is not constrained: for each row in the table the optimal set was found. The fact that, in this and many other cases, the optimal set of cutoffs did not change from one row to the next shows that the choice of the optimal set is not very sensitive to these changes (ranging up to 55%) in relative workloads.

The results for Saturday at 5 AM show a typical pattern. As successively more calls are shifted from priority 2 to priority 1, the expected wait in both priority 1 and priority 2 increases. The increase for priority 1 is due to the increased number of calls given top priority, and the increase for priority 2

Table V. Shifts of Calls from Priority 2 to Priority 1 (Peoria Data)[a]

Day	Time	Total calls for service	Expected total cost	Optimal cutoff			E(W1)	E(W2)	E(W3)	% moved
				N_1	N_2	N_3				
Sat	5	9.8	36.150	9	8	8	1.70	8.42	81.70	0
Sat	5	9.8	38.085	9	8	8	1.77	8.49	79.59	10
Sat	5	9.8	41.279	9	8	8	1.89	8.60	76.56	25
Sat	5	9.8	43.613	9	8	8	1.97	8.68	74.62	35
Sat	5	9.8	46.121	9	8	8	2.05	8.76	72.75	45
Sat	5	9.8	48.809	9	8	8	2.13	8.85	70.94	55

[a]Total calls for service is aggregate arrival rate of calls per hour. The E(W) are expected waits in minutes.

is due to the fact that those remaining in that category now have more people ahead of them. What is not obvious is why the expected wait for priority 3 callers should diminish. The answer lies in the fact that as more priority 2 calls are moved to priority 1, there is more complete utilization of servers: because the cutoff for priority 2 is less than the cutoff for priority 1, there will be times when a priority 2 caller waits for service even though a server is available, and there will be less of this kind of waiting when there are fewer calls in category 2. That is, the entire set of priority 1 and priority 2 calls, taken as a group, will be served faster if some of the priority 2 callers are moved up, and this will allow the servers to attend to category 3 callers sooner. In the extreme case, we could simply eliminate the distinction between category 1 and category 2, thus speeding up service to category 3.

There is an important implication here that can be generalized to more than three categories. Obviously, it is in the interests of priority 3 callers to have those ahead of them served as quickly as possible. Because dividing those ahead of them into two groups results in more cautious use of response units (they are not dispatched to group 2 calls unless there are more than $N-N_2$ cars free), group 3 (and, in general, any group) would be better off if all of those ahead of them are put into a single priority category. Obviously, the disadvantage of doing that is the increased expected wait for top-category calls. It is because the costs of a minute's delay are not the same for categories 1 and 2 that they should not be merged.

In another set of simulations we moved varying percentages of calls from priority 2 to priority 3. The effect of this change is that the expected wait diminishes for all three categories, as increasing percentages of the calls in priority 2 are moved to priority 3. This is easily understandable for priorities 2 and 3: calls remaining in category 2 are now ahead, not only of those which had been behind them, but also of some which had been equal to them in priority; for those in priority 3, while the number of calls in their category increases, the total number of calls in the system does not change, and there now are fewer calls which are unequivocally ahead of them. But it is not so easy to see why the expected wait should fall for priority one calls. The explanation lies in the fact that, by moving some calls down from priority 2 to priority 3 we are effectively increasing the "reserve" held for the possible arrival of priority one calls. In the extreme case, if the cutoffs are, say, 14/13/12, if we were to move all of the priority 2 calls to priority 3, we would effectively be holding two cars, rather than one, ready for possible priority 1 callers; thus, two nearly simultaneous arrivals of priority 1 calls could be handled with no delay for either of them. If we move down fewer than all the priority 2 calls, the effect is less, but we are still reducing the likelihood that a priority 1 caller will have to wait for service while a priority 2 caller is being served.

In a number of simulations we recognized that the probability of requesting two servers is not the same in all priority categories. In these cases, in order to observe the consequences of a pure change in relative workloads, we had to be careful not only to maintain a constant total number of callers, but also to maintain a constant total number of servers requested. That is, we adjusted the probabilities of requesting two servers in a way that leaves the overall workload on the system unchanged.

Making these adjustments, simulations were run in which various percentages of priority 1 callers were moved to priority 2 at different hours of the week. Not surprisingly, we found that the expected wait for priority 1 callers diminishes as first 10% and then 25% of priority 1 callers are shifted to priority 2. What needs some explanation is why the priority 1 wait *increases* when we move even more people (40, 55, 70, or 80%) out of category 1. The answer is to be found in the underlying optimization process: with a substantial number of callers moved to priority 2, it becomes optimal to allocate a greater proportion of available resources to serving priority 2 callers, that is, the priority 2 cutoff goes up.

Table VI shows one case where, as more and more calls are moved from priority 1 to priority 2, the optimal set of cutoffs changes from 7/6/5 to 7/7/5 (when 40% of the priority 1 calls are moved); that is, as priority 2 becomes relatively more important (as measured by number of calls), the optimization procedure allocates it more resources. That is why the expected wait for priority 1 callers rises. The expected wait for priority 2 callers increases until their cutoff improves, and then as more calls are put into category 2 their expected wait again worsens.

It is interesting to note that the expected wait for priority 3 callers changes along with that of priority 2 callers: it worsens as up to 25% of the priority 1 calls are moved to priority 2 and improves when an extra vehicle

Table VI. Shifts of Calls from Priority 1 to Priority 2 (Adjusted for Different Probabilities of Requesting Two Servers; Peoria Data)[a]

Day	Time	Total calls for service	Expected total cost	Optimal cutoff			E(W1)	E(W2)	E(W3)	% moved
				N_1	N_2	N_3				
Wed	8	7.6	4.002	7	6	5	0.25	1.09	7.77	0
Wed	8	7.6	3.678	7	6	5	0.23	1.12	7.86	10
Wed	8	7.6	3.252	7	6	5	0.21	1.17	8.00	25
Wed	8	7.6	2.820	7	7	5	0.28	0.30	7.30	40
Wed	8	7.6	2.358	7	7	5	0.27	0.30	7.30	55
Wed	8	7.6	1.920	7	7	5	0.26	0.30	7.30	70
Wed	8	7.6	1.614	7	7	6	0.43	0.50	2.70	80

[a]The E(W) are expected waits in minutes.

is made available to priority 2, thus in effect reducing the backlog that must be served before priority 3 calls can be serviced. When fully 80% of the priority 1 calls are demoted to priority 2, the total importance of giving preference to priority 1 callers diminishes and the optimization process allocates an additional vehicle to priority 3 calls, which, although no more numerous in an absolute sense, are then relatively more important. At that point their expected wait drops sharply (from 7.3 to 2.7 min).

5. SUMMARY

In this article we examine a differential response strategy, which means the following.

- Every caller is assigned to a priority category.
- In some circumstances response will be deliberately delayed, even though some response units are available for dispatch. More formally, arriving priority i callers are queued whenever the number of busy patrol units equals or exceeds the priority i cutoff number.
- The number of response units dispatched to a single call is variable.
- In some cases there will be no response to a call (the caller will be referred to an alternative system). More formally, it may be that $N_i = 0$ for some i.

Using data for police calls in Hartford, Connecticut, and Peoria, Illinois, we have shown that this strategy can substantially reduce the total expected cost of delays in a police response system, thus more efficiently allocating resources than a simple first-come, first-served queue. We have also shown that our gradient search method will find the optimal set of cutoffs with many fewer iterations of the Schaack and Larson model than would be required by complete enumeration. We use the gradient method to investigate how the optimal cutoffs vary in response to changes in several important variables.

First, we showed that it takes increasingly large changes in cost to cause changes in the optimal set of cutoffs. This means that within certain ranges, the optimal cutoffs are not very sensitive to the, necessarily subjective, choice of what cost to assign to a minute's wait for each priority category.

Second, we ascertained that increasing the workload on the system (the total number of calls per hour) changes the differentials among the cutoffs in the optimal set. As the workload increases from very light to very heavy, the differences between the cutoffs increase and then decrease: at both very low and very high demand there is little or no advantage to a system of differential cutoffs.

Third, simulations in which we changed the relative frequency of calls from the various priority classes (without changing the total number of calls) show that the optimal set of cutoffs is not very sensitive to such changes if we assume that all priorities have the same probability of requesting multiple servers. The results are not always what we would expect, but in each case they can be explained. It is not surprising that shifting priority 2 calls to priority 1 increases the expected wait for both of those groups, but it is not obvious why this should decrease the priority 3 wait; the answer is that this decrease derives from the less cautious use of available response units. This suggests that, in general, it is in the interest of any group that the groups above it be consolidated. Shifting calls from priority 2 to priority 3 causes the expected wait to decrease for all three groups, except when the optimal set of cutoffs changes.

Our results indicate that the optimal cutoffs are more sensitive to changes in relative workloads in situations where the probability of needing two cars varies across priority groups. Perhaps the most interesting result is that if we allow the cutoffs to adjust optimally, reducing the number of priority 1 calls will not necessarily reduce the expected waiting time for the remaining priority 1 calls. That is, rewriting the rules to reduce the number of calls that qualify as top priority may not lead to faster response for those who do qualify; this is because changes in the optimal set of cutoffs may allocate more resources to lower priority calls.

The practical implication of our research is that in many situations a very simple differential response strategy, one that holds back one or two response units for dispatch only to top priority calls, will result in more efficient allocation of police resources. A more sophisticated differential response strategy, one that uses a computer algorithm to calculate the optimal number of units to hold back for each priority category, can be expected to improve efficiency even more. We would be happy to discuss use of our computer model with others.

At present, work is under way to implement this strategy in Hartford. In order to get the best cutoffs, we are redoing our calculations using data that have two advantages over those used for this article: we have separate data for each of two zones, and for each hour of the week we are using data that are the average for that hour over 4 weeks. Also, we plan to calculate a separate set of cutoffs for each season of the year. We believe that the differential response strategy based on these cutoffs will significantly reduce the total cost of delays in Hartford.

REFERENCES

Bottoms, A. M., Nillson, E. K., and Olson, D. G. (1972). *Allocations of Resources in the Chicago Police Department*, U.S. 2700-0151, Government Printing Office, Washington, DC.

Cahn, M. F., and Tien, J. M. (1981). *An Alternative Approach to Police Response, Wilimington Management of Demand Program*, U.S. Department of Justice, National Institute of Justice, Washington, DC.

Chaiken, J. M. (1978). Transfer of emergency service deployment models to operating agencies. *Manage. Sci.* 24: 719–731.

Chaiken, J. M., and Dormont, P. (1975). *Patrol Car Allocation Model: Executive Summary, User's Manual, and Program Description*, R-1786/1,2,3-HUD/DOJ, Rand Corp., Santa Monica, CA.

Chaiken, J. M., and Dormont, P. (1978). A patrol car allocation model: Background, capabilities and algorithms. *Manage. Sci.* 24: 1280–1300.

Chelst, K. R., and Barlach, Z. (1981). Multiple unit dispatches in emergency services. *Manage. Sci.* 27: 1390–1409.

Green, L. (1984). Multiple dispatch queueing model of police patrol operations. *Manage. Sci.* 30: 653–664.

Green, L., and Kolesar, P. (1989). Testing the validity of a queueing model of police patrol. *Manage. Sci.* 35: 127–148.

Larson, R. C. (1972a). Improving the effectiveness of New York City's 911. In Drake, A. W., Keeney, R. L., and Morse, P. M. (eds.), *Analysis of Public Systems*, MIT Press, Cambridge, MA.

Larson, R. C. (1972b). *Urban Police Patrol Analysis*, MIT Press, Cambridge, MA.

Larson, R. C. (1974a). Illustrative police sector redesign in District 4 in Boston. *Urban Anal.* 2(1): 51–91.

Larson, R. C. (1974b). A hypercube queueing modeling for facility location and redistricting in urban emergency services. *J. Comput. Operat. Res.* 1(1): 67–95.

Larson, R. C. (1975). Approximating the performance of urban emergency service systems. *Operat. Res.* 23(5): 845–868.

Larson, R. C. (1989). The new crime stoppers. *Technol. Rev.* Nov.–Dec.: 26–31.

Larson, R. C. (1990). Rapid response and community policing: Are they really in conflict. Series Paper, National Center for Criminal Justice, The Michigan State University, East Lansing.

Larson, R. C., and Chaiken, J. (1972). Methods for allocating urban emergency units: A survey. *Manage. Sci.* 19(4): 110–132.

Larson, R. C., and Franck, E. (1978). Evaluating dispatching consequences of automatic vehicle location in emergency services. *J. Comput. Operat. Res.* 5: 11–30.

Larson, R. C., and McEwen, T. (1974). Patrol planning in the Rotterdam Police Department. *J. Crim. Just.* 2(3): 235–238.

Larson, R. C., and McKnew, M. A. (1982). Police patrol-initiated activities within a system queueing model. *Manage. Sci.* 28(7): 759–774.

McEwen, J. T., Connors, E. F., III, and Cohen, M. I. (1986). Evaluation of the Differential Police Response Field Test (Executive Summary and Research Report), U.S. Department of Justice, National Institute of Justice, Washington, DC.

Schaack, C., and Larson, R. C. (1986). An N server cutoff priority queue. *Operat. Res.* 34(2): 257–266.

Schaack, C., and Larson, R. C. (1989). An N server cutoff priority queue where arriving customers request a random number of servers. *Manage. Sci.* 35(5): 614–634.